Studying Lyon's Vélo'V: A Statistical Cyclic Model

**Pierre Borgnat · Patrice Abry ·
Patrick Flandrin**

Abstract Lyon's community bicycle program called *Vélo'v* is a major initiative in shared public transportation, in activity since May 2005. It is studied here at a global level, to assess the evolution with time of the number of hired bikes. Based on the entire *Vélo'v* data set, up to December 2007, a statistical model is proposed to describe the daily and weekly patterns in a cyclostationary manner, jointly with the non-stationary evolutions over larger time-scales larger. Combining this model with linear statistical regression, a procedure is developed for the prediction of the number of bikes hired per hour. This prediction method involves several explanation factors such as the number of subscribed users, the time in the week, the occurrence of holidays or strikes, and weather parameters (temperature, volume of rain). The conclusion is that, for most days, the observation of the number of actually hired bicycles is satisfyingly explained and predicted by the model proposed here.

1 Lyon's community shared bicycles: *Vélo'v*

Vélo'v is a program of community shared bicycles that is deployed in the cities of Lyon and Villeurbanne [1]. It is one of the leading program of this kind in France, running actively by JCDecaux from May 2005 on. It consists now of 4000 available bicycles that can be hired at any of the 334 stations, spread all over the towns, and returned back later at any station. With the increasing need of green and versatile public transportation in cities, there is a current development of such community shared equipement programs in many european cities, e.g., the *Vélib'* program in Paris since July 2007, or the *Bicing* program launched in Barcelona in March 2008 [2]. The specificity of those programs, as compared to old-fashioned rental systems, is that stations can be accessed freely at any time and are fully automatized and computerized. This makes possible a global management and potentially some real-time survey of the state of the system.

P. Borgnat (Corresponding Author), P. Abry, P. Flandrin
CNRS, Université de Lyon, ENS Lyon, Laboratoire de Physique (UMR 5672), France
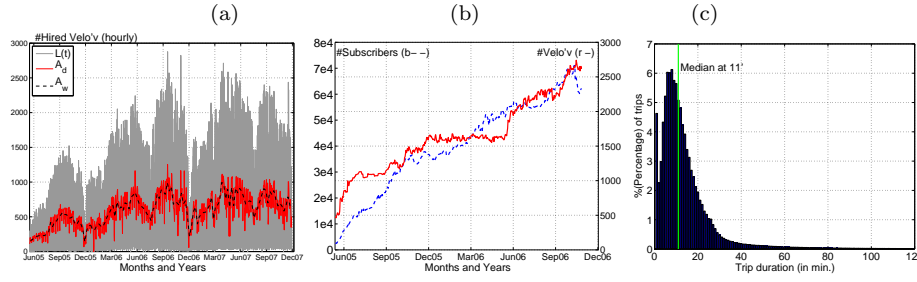Tel.: +33-47272-8691— Fax: +33-47272-8080 — E-mail: {firtsname.name}@ens-lyon.fr

Fig. 1 Global features of *Vélo'v*. (a) Number of bikes hired per hour, $L(t)$, and its average per day A_{d} and per week A_{w}. (b) Evolutions along time of the numbers of subscribers N_s of the program (dashed line, in blue) and of available bikes N_v (solid line, in red). (c) Rental duration distribution (in %).

Currently, automatized station reports are collected into a central data set and mostly used a posteriori. Yet, there is a strong incentive to evolve toward less empirical management of the system, for instance by being able to increase or redeploy in real-time the available bicycles if need be at a specific time one day. Therefore, a challenge consists of quantitatively studying the behaviour of the system, so as to be able, in the future, to regulate and optimize it (e.g., increase bicycle availability). The present work proposes a global and statistical study of the evolution of the number of bicycles hired along time. Objectives are not only to identify its temporal patterns but, going way further than [2] in the modeling, to propose a statistical model for the temporal patterns that encompass their cyclostationarity and their non-stationarity. Finally, this model is used to predict the number of bicycle rentals on a daily or hourly basis.

Anonymized data were made available to us by JCDecaux and the Grand Lyon. The data set consists of the records of all bike rental operations (starting and ending points and times), over more than two years of exploitation. Companion studies are made from this data set about the specific spatial patterns that can be exhibited in journeys [3]. Here, the focus will be on aggregated data over the whole network, and its evolution along time, as shown in Fig. 1 (a).

In Fig. 1 (b), the number of available bikes and stations and the number of subscribed users (note that bikes can also be used without subscription, with short-term registration cards) are shown. This sketches the progressive deployment and the increase in popularity of the program.

2 Empirical patterns of the number of rentals

A first question is to choose a time scale Δ to aggregate the number of new rentals. The trade-off here is usual: the smaller Δ is, the larger the fluctuations are, whereas a larger Δ may smooth the signal with the risk of losing some relevant temporal feature. Fig. 1 (c) displays the distribution of rental duration. This distribution is large, yet there is a mode and the median of this duration, equal to 11 minutes, is representative of it. Choosing $\Delta = 1h$ is enough to smooth out some of the differences of individual rentals, while keeping the global evolution of their collection.

On Fig. 1 (a) is shown the number of hired bikes, aggregated on Δ, from May 2005 to Dec. 2007, superimposed with the per hour hiring numbers averaged over the day and

over the week. Two features are striking. First, the mean is non-stationary and evolves with time. The first interpretation is to be related to the increase of the size and of the popularity of the program, as already seen in Fig. 1 (b). A complementary explanation is that the use of *Vélo'v* also depends on the season (with less users during winter, or during holidays). The second feature is a strong modulation of the hirings with the moment in the week: this cyclic evolution (more properly called cyclostationarity) comes from the evident fact that from a social point of view, days and hours are not equivalent for people. Those two features, non-stationarity and cyclostationarity, are those the model proposed below aims at accounting for.

3 Model for the cyclic temporal patterns

Let us first study non-stationary patterns on time scales larger than the day. Empirically, we put them in evidence by computing, from the hourly hirings $L(t)$, the mean $A_d(d)$ over each day (d is the variable of day). Then, inspired from cyclostationary methodologies [5], we estimate with a cyclic mean over the week, the weekly temporal pattern for $L(t)$ (displayed on Fig. 2 (a)):

$$< L(t) >_c = \frac{1}{N_w} \sum_{d=0}^{N_w-1} L(t + 168\,d), \quad \text{for} \quad t = 0\,..\,(24 \cdot 7 - 1)\text{h} = 0\,..\,167\text{h}. \quad (1)$$

Here, N_w is the number of weeks of data used. The result shows patterns also observed in the Barcelona program [2]. During week-days, three peaks are seen: in the morning (8am-9am), noon (12am-1pm) and end of afternoon (6pm-7pm, this one being the biggest). During week-ends, the pattern changes, with mostly a large peak spread during the afternoon, having a maximum around 5pm (with only a small increase on top of that at noon). These features match intuitive interpretations about the fact that people use bicycle transportations mostly during the day to go to and come back from work, or during lunch break, whereas during the week-end, the major trend is to take an afternoon pleasure ride.

From the cyclic means, let us write $A_{\text{mod}}(d_7) = \sum_{(d_7)} < L(t) >_c$ the average number of rentals per day d_7, that depends only on the position of the day in the week (and we write d_7 to represent the day in the week). For a quantitative approach about the temporal pattern, we propose the following model:

$$L(t) = L_{\text{mod}}(t) + F(t) = A_d(d) \frac{< L(t) >_c}{A_{\text{mod}}(d_7)} + F(t), \quad (2)$$

where $F(t)$ is a fluctuation not accounted for by the cyclic model. In Fig. 2 (b), we illustrate the model for a specific range of days, to show that it usually holds well, even when specific occasions change the flow of days, such as holidays or festivities (here we illustrate that on December, the 8th, which is a specific festivity day in Lyon). It has not yet been discussed so far how to predict of the fluctuations F and the amplitude A_d. This is the objective of the next Section.

4 Statistical forecasting of the number of bikes hired

Let us now turn to the prediction of the evolution of the per hour number of bikes rented taking into account factors that are external to the cyclic pattern. Using the
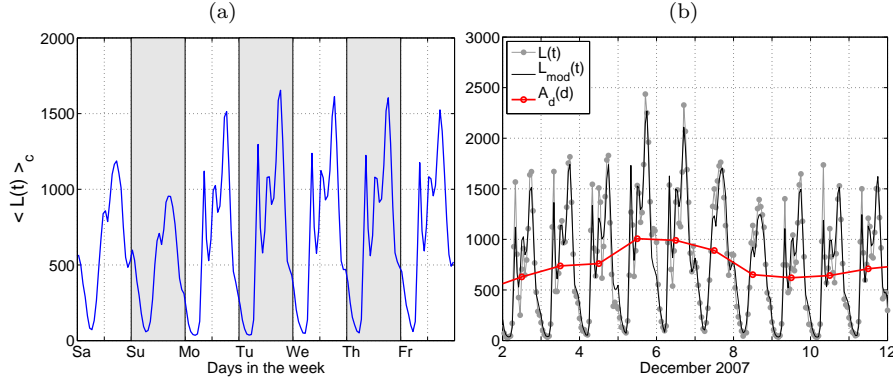
Fig. 2 Cyclic models and comparison to data. (a) Model $< L(t) >_c$ giving the typical expected evolution over the week. (b) Examples of $L(t)$ for some chosen days, compared to the model $L_{\mathrm{mod}}(t) + \widehat{F(t)}$. Here, we show the days around December, the 8th, in Lyon (which was a Saturday in 2007), showing that the model holds qualitatively well here.

model in eq. (2), prediction is split into two subparts: First, the prediction of the non-stationary amplitude $A_{\mathrm{d}}(d)$ for a given day; Second, the prediction of the fluctuations $F(t)$ at a specific hour. The corresponding time scales being different, it is sound to try to predict them separately.

Prediction of $A_{\mathrm{d}}(d)$. Our starting proposition is that the factors explaining $A_{\mathrm{d}}(d)$ are to find among the following ones: (i) the weather conditions summarized by the average temperature $T(d)$ over one day and the volume of rain $R(d)$ (in mm); to remove the mean from the temperature, we will use $\delta T(d) = T(d) - < T(d) >$ (in $^{\mathrm{o}}$C); (ii) the development and popularity of the program: the number of subscribed users $N_s(d)$, the number of bicycles available $N_v(d)$; here again, we take deviations $\delta N_s(d)$ and $\delta N_v(d)$ between the real value and the value at the end of the data (XII.2007) where the system is supposed to have reached its final state; (iii) specific conditions such as holidays, with a marker $J_h(d)$ taking value 0 usually and 1 for those specific days, or strikes with marker $J_s(d)$. A linear regression model is written as:

$$\widehat{A_{\mathrm{d}}(d)} = \alpha_0(d_7) + \alpha_1 \delta N_s(d) + \alpha_2 \delta N_v(d) + \alpha_3 \delta T(d) + \alpha_4 R(d) + \alpha_5 J_h(d) + \alpha_6 J_s(d), \quad (3)$$

where features δN_s to R have been normalized to variance 1, and where the term $\alpha_0(d_7)$ describes the mean with an additional linear dependence on the position of day d_7 during the week (from Monday to Sunday):

$$\alpha_0(d_7) = A_0 + c_1 \left(A_{\mathrm{mod}}(d_7) - < A_{\mathrm{mod}}(d_7) >_{d_7} \right). \quad (4)$$

This dependence is needed because, as seen on Fig. 2 (a), the expected number of hirings each day varies from Monday to Sunday; it is, for instance, smaller during the week-ends. Solving this problem of linear regression using standard least square minimization, we obtain the results reported in Table 1. Confidence intervals are reported along with the estimated values of the coefficients because, even though computed under Gaussian hypothesis, which does not hold for many factors, it assists us in the interpretation of the relevance and importance of each factor. Results call for the following comments. First, the term depending on the day $\alpha_0(d_7)$ is simple enough: it

Table 1 Statistical model for $A_{\mathrm{d}}(d)$ as per eq. (3). For the different linear coefficients associated to the factors in play, we report the estimated value (est.) and its Confidence Interval (under Gaussian assumption), given by $[\mathrm{CI}_-, \mathrm{CI}_+]$.

Factor	$\alpha_0(d_7)$		$\delta N_s(d)$	$\delta N_v(d)$	$\delta T(d)$	$R(d)$	$J_h(d)$	$J_s(d)$
coeff.	A_0	c_1	α_1	α_2	α_3	α_4	α_5	α_6
est.	17 370	1.05	1 860	-120	2270	-1280	-2900	20
CI_-	17 050	0.90	1 210	-720	1980	-1520	-3700	-2900
CI_+	17 680	1.18	2 560	+490	2560	-1030	-2100	+2900

consists of a constant A_0 whose value is close to the average number of hired bikes per day during the last months in the data set (17 500 during the last 4 months of 2007), with a linear correction (with factor close to 1) that takes into account the dependence with the day in the week. Second, some factors play an important role in controlling $A_{\mathrm{d}}(d)$: a larger number of subscribers increases it; weather factors act in an expected manner: the warmer, the larger the number of bikes used (and conversely), under heavy rain, $A_{\mathrm{d}}(d)$ decreases. The factor pertaining to holidays J_h also impacts $A_{\mathrm{d}}(d)$: there is a decrease (whose relevance is assessed by the confidence interval) during holidays—a feature that appears qualitatively in Fig. 1 (a) and is explained by the fact that people are out of town during holidays. The two remaining factors show non conclusive effects; the number of available bikes does not impact much $A_{\mathrm{d}}(d)$ and this can interpreted by looking again at Fig. 1 (b): the numbers of subscribers and the number of bikes follow roughly the same evolution at the beginning, then are constant. This lack of influence hence results from the fact that a part of the evolution is already accounted for by the evolution of N_s, and by the fact that there seems to be no major depletion of bikes as confronted to subscribers. Finally, strikes are a non conclusive factor, mostly because of the scarce number of such events in the current data set.

Using this linear regression model, it has been possible to predict the amplitude of the number of bikes rented per day, depending on all the external factors proposed here. If one would use only the average number of hirings $A_{\mathrm{mod}}(d_7)$ adjusted only for the position of the day d_7 in the week and without the other non-stationary factors, the rms of the error value between observed data $A_{\mathrm{d}}(d)$ and this number, divided by the mean of this amplitude, would conduct to 30% of mean relative error. Using the model $\widehat{A_{\mathrm{d}}(d)}$, it decreases to 12%. Clearly there is still room for improvement, yet the quantitative gain is non negligible and, more importantly, the interpretation of the dependence with the various factor shows their relevance.

Prediction of fluctuations and its anomalies. Let us know turn to the fluctuation term $F(t)$, whose standard deviation is 210 (in bikes hired per hour; it can be compared to the mean of $L(t)$ equal to 655 hired bikes per hour). A standard empirical spectrum analysis shows that is is well modeled by a parametric AR(1) process:

$$F(t) = a_1 F(t-1) + I(t), \tag{5}$$

where $a_1 \simeq 0.81$ and $I(t)$ is a white innovation of standard deviation equals to 120. The coefficient a_1 and the order of the model were estimated using the classical Levinson-Durbin algorithm [6]. This leads to a general prediction scheme for the number of hourly rentals that follows eq. (2) with $\widehat{A_{\mathrm{d}}(d)}$ obtained from Eq. (3) and $\widehat{F(t)} = a_1(L(t-1) - L_{\mathrm{mod}}(t-1))$. In Fig. 2 (b), the displayed model is built using these estimates $\widehat{A_{\mathrm{d}}(d)}$ and

$\widehat{F(t)}$ and eq. (2). It works satisfactorily in following the observed variations of $L(t)$ along time. Using this improved scheme including prediction of the fluctuations, the standard deviation of the error of the global prediction decreases from 210 bikes to 120 bikes per hour, i.e., the std. of the innovation I, which, by nature of the approach, cannot be predicted. The improvement, which is clear, is not larger because of many anomalies existing in the data that are not explained by the factors used here. We propose that, for instance, taking into account the rain at the time scale of the hour (rain being often an event that does not last several hours), would improve the prediction of the fluctuation from one time to the next one. This is not done here.

5 Conclusion and on-going work

This study is a preliminary step toward a better and quantitative understanding of the usage of the *Vélo'v* program in Lyon. The combination of non-stationarity and cyclo-stationarity that drives the number of hourly hirings was modeled here using suitable statistical signal processing methods. A first result is to confirm that the temporal pattern over the week that was already exhibited in studies on other bicyle sharing programs seems to hold in different cities (when corrected for non-stationary evolu-tions and short-lived anomalies). A new result is the possibility to model the number of per hour rental along time and predict its evolution based both on characteristics of the deployed program (number of bikes and or subscribers), and on external conditions such as weather. Combined together, they explain most of the remaining deviations from the model.

Spatial patterns are studied with complementary techniques in a joint work [3]. The objective is that these quantitative studies of the data set of all trips with *Vélo'v* could help on the one hand sociological studies on community shared transportation systems and, on the other hand, managers that have to start or run such programs.

6 Acknowledgments

References

1. http://www.velov.grandlyon.com/
2. Froehlich, J., Neumann, J., and Oliver, N. "Measuring the Pulse of the City through Shared Bicycle Programs" *International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems - UrbanSense08*, Raleigh, North Carolina, USA (November 4, 2008).
3. Borgnat, P., Fleury, E., Robardet, C., and Scherrer, A, "Spatial analysis of dynamic move-ments of Vélo'v, Lyon's shared bicycle program", Submitted Preprint (March, 2009).
4. Girardin, F. "Revealing Paris Through Velib' Data" http://liftlab.com/think/fabien/2008/02/27/revealing-paris-through-velib-data/ (2008).
5. Gardner, W., Napolitano, A., and Paura, L., "Cyclostationarity: Half a century of research", *Signal Processing* vol. 86 (4), p. 639–697 (2006).
6. Priestley, M.B. *Spectral analysis and times series.* Academic Press, San Diego (1981).